제33회 한국정보처리학회 춘계학술발표대회 논문집 제17권 제1호 (2010. 4)

SVD를 이용한 블로그 공간 분석

김기남, 김상욱
한양대학교 전자컴퓨터통신공학과
e-mail: kinam@zion.hanyang.ac.kr

Analysis of a Blogosphere using SVD

Ki-Nam Kim, Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University

요 약

본 논문은 행렬 분해 기법인 SVD를 이용하여 블로그 공간을 분석한다. 분석 결과 블로그 공간에서 커뮤니티들을 발견했고, 각 커뮤니티에 속한 영향력 있는 블로그와 포스트를 발견했다.

1. 서론

사회 연결망은 사회 안에 존재하는 사람들 간의 관계를 네트워크로 표현하는 하나의 방법이다[1]. 최근 들어, 인터넷의 발전과 함께 사회 연결망은 온라인상에서도 나타나고 있다. 블로그 공간은 온라인 사회 연결망의 대표적인 예이다. 블로그는 블로그의 주인인 블로거가 자신의 의견이나 생각을 글로써 온라인상에 저장할 수 있는 일종의 개인 홈페이지이다. 블로그 안에 저장되어 있는 이러한 글을 포스트라고 한다. 블로거는 다른 블로그에서 자신이 관심을 가지는 포스트를 대상으로 여러 가지 행동을 할 수 있는데, 이러한 행동들로는 스크랩, 덧글 등이 있다. 스크랩은 블로거가 다른 블로그에 존재하는 포스트를 복사해서 자신의 블로그로 가지고 오는 행동을 말하며, 덧글은 포스트에 자신의 생각을 글로써 남기는 행동을 말한다.

블로그 공간 규모가 점점 증가하면서 블로그 공간을 분석하려는 연구가 활발하게 진행되고 있다[2]. 본 논문에서는 블로그 공간을 블로그와 포스트의 차원을 가지는 행렬(matrix)로 표현하고, 기존의 행렬 분해 기법인 SVD(singular value decomposition)[3]를 이용하여 블로그 공간을 분석하고자 한다.

2. SVD

SVD는 행렬 분해 기법으로 고유치(eigenvalue)와 고유벡터(eigenvector)에 기반을 두고 있다. 적용되는 분야는 주성분 분석, 이미지 패턴 인식, 행렬 압축 등이 있다[3]. SVD의 개념은 다음과 같이 기하학적으로 설명할 수 있다. $n \times m$ 행렬 A는 m개의 차원을 가지는 공간에 n개의 점으로 생각할 수 있다. SVD는 해당 공간에 존재하는 점들과 최소 제곱 오차(Least square error)를 최소화 하는 r개의 직교벡터(orthogonal vector)를 찾는 방법이다. 즉, SVD는 해당 공간에 존재하는 n개의 점을 잘 표현하는 r개의 직교벡터를 찾는 것을 의미한다.

SVD를 통해 $n \times m$ 행렬 A는 (식 1)과 같이 세 개의 행렬의 곱으로 분해된다. (식 1)에서 행렬 곱 연산은 벡터 곱 연산의 합으로도 나타낼 수 있다(spectral decomposition).

$$\begin{aligned} A &= U \times \Lambda \times V^t \\ &= \lambda_1 u_1 \times v_1^t + \lambda_2 u_2 \times v_2^t + ... + \lambda_r u_r \times v_r^t \\ &= \sum_{i=1}^{r} \lambda_i u_i \times v_i^t \end{aligned} \quad \text{(식 1)}$$

이때 행렬 U, Λ, V는 각각 $n \times r$, $r \times r$, $m \times r$ 크기를 가진다. r은 행렬 A의 랭크(rank)이며, 랭크는 행렬을 구성하는 벡터(열벡터, 행벡터) 중에서 서로 독립인 벡터(independent vector)들의 수이다. 행렬 U와 V은 직교행렬(orthogonal matrix)이고, 행렬 Λ는 대각행렬(diagonal matrix)이다. Λ의 대각 요소에는 A의 고유치의 제곱 값이 부여되고, 그 외의 요소에는 0이 부여된다. 벡터 u_i와 v_i는 행렬 U와 V의 열백터(column vector) 이고, 이를 특이벡터(singular vector)라 한다. λ_i는 Λ 행렬의 대각요소이고, 이를 특이값(singular value)이라 한다. 특이값 ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$)은 양수이고 내림차순으로 정렬된다.

3. 블로그 공간에서 SVD

본 논문에서는 블로그 공간에서 특정 주제에 관심을 가지는 블로그 집단과 특정 주제를 나타내는 포스트 집단, 이러한 두 집단을 커뮤니티라 정의한다[4]. 블로그 공간에서 '가' 주제에 관심을 가지는 블로그 집단과 '나' 주제에 관심을 가지는 블로그 집단이 있을 때, 일반적으로 이러한 두 블로그 집단이 스크랩하는 포스트 집단들은 서로 다르다. 즉, 관심 주제에 따라서 블로그 집단은 다른 스크랩 패턴으로 포스트 집단을 스크랩한다. 다른 스크랩 패턴은 블로그 공간에서 직교벡터로 나타나며, SVD는 직교벡터를 찾는 과정에서 블로그 공간에 존재하는 커뮤니티들을 추출한다. SVD를 통해 얻어진 랭크는 커뮤니티들의 수를 나타내고, 특이값 λ_i은 분해된 커뮤니티가 블로그 공간에서 차지하는 비중을 나타낸다. 특이벡터 u_i는 전체 블로그들과 커뮤니티 i의 주제와의 관련 정도, v_i는 전체 포스트들과 커뮤니티 i의 주제와의 관련 정도를 점수로 나타낸다. 이 점수를 이용하여 특정 커뮤니티의 주제에 영향력

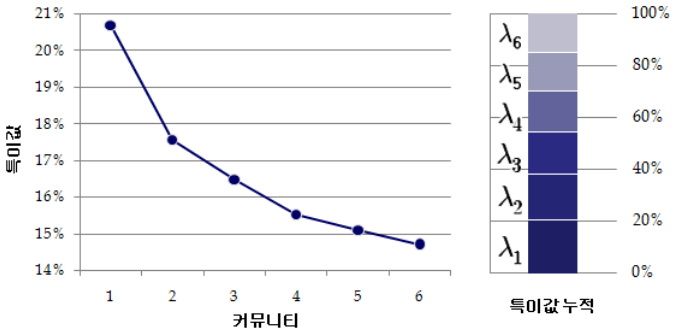
있는 블로그와 포스트를 발견할 수 있다.

4. 실험 및 결과

실험에서 사용한 데이터는 국내 블로그 사이트로부터 2006년 4월부터 수개월간 수집하여 익명으로 처리한 데이터이다. 거의 활동하지 않는 블로그와 거의 스크랩을 받지 않은 포스트는 실험에서 제거한다.

SVD를 이용해서 블로그 공간을 분석하기 위해서 본 논문에서는 다음과 같이 블로그 공간을 $N \times M$ 행렬 A로 변환한다. N은 블로그의 수이고 M은 포스트의 수이다. 블로그 i가 포스트 j 사이의 스크랩이 존재할 경우, 행렬 A의 요소인 A_{ij}는 1의 값을 갖고, 블로그 i와 포스트 j 사이의 스크랩이 존재하지 않을 경우, A_{ij}는 0의 값을 갖는다.

SVD를 이용하여 블로그 공간을 분석한 결과는 다음과 같다. 분해된 행렬의 랭크는 6이다. (그림 1)의 (a)는 6개의 커뮤니티 i와 각 커뮤니티에서 특이값 λ_i을 보여준다. 특이값 λ_1은 첫 번째 커뮤니티가 블로그 공간에서 약 21%의 비중을 차지하는 것을 나타낸다. (b)는 6개의 특이값의 합이 블로그 공간에서 100%비중을 차지하는 것을 보이며, 이 결과는 블로그 공간이 6개의 큰 커뮤니티로 이루어져 있음을 나타낸다.



(a) 각 커뮤니티의 특이값 (b)특이값 누적 막대

(그림 1) 특이값 관찰.

커뮤니티 주제 분석은 SVD를 통해 도출된 커뮤니티의 주제를 알기 위해서 필요하다. 일반적으로 커뮤니티에 속한 포스트들의 제목에 빈번히 등장하는 키워드가 있다면 해당 커뮤니티의 주제는 이 키워드와 관련이 높을 것이다. 따라서 본 논문에서는 특정 커뮤니티에 속한 포스트들의 제목에서 빈번이 발생하는 키워드를 추출하여 해당 커뮤니티의 주제를 알아낸다. <표 1>은 6개의 커뮤니티에 속하는 포스트들에서 빈번히 나타나는 키워드와 이 키워드로부터 도출한 커뮤니티의 주제를 나타낸다.

특이벡터 u_i와 v_i를 점수를 이용하여 커뮤니티 i의 주제와 관련이 높은 블로그와 포스트를 찾을 수 있다. <표 2>는 첫 번째 커뮤니티의 주제와 관련성이 높은 상위 포스트 5개의 제목, 점수, 스크랩 수를 나타낸다. <표 2>에서 알 수 있듯이 상위 5개의 포스트는 모두 첫 번째 커뮤니티의 주제인 '요리'에 관련된 내용을 담고 있다. 또한 확인 결과 실제로 주제에 관심을 가지는 블로그들에게 많은 스크랩을 받은 영향력 있는 포스트이다.

<표 1> SVD 방법을 이용해서 추출한 커뮤니티의 주제들

커뮤니티	주제	빈번히 등장하는 키워드
1	요리	요리, 볶음, 맛, 치즈
2	만들기	리폼, 박스, 사과상자, 만들기
3	연예	스타, 포토, 영화, 연예
4	인테리어	인테리어, 라벨, 리폼, 모음
5	팝 음악	사랑, 슈퍼주니어, OST, 동방신기
6	클래식 음악	악장, 소나타, 협주곡, 바흐

<표 2> 첫 번째 커뮤니티에서 영향력 있는 포스트 5개

상위 포스트 제목	점수	스크랩
한입에 쏙쏙쏙~~ 한입떡꼬치..	0.154	857
새콤~달콤~매컴.. 입맛없다 쫄쫄이 굶지말고 쫄면즐기자~	0.151	808
아직도 피자를 시켜드세요? 초간단 피자~ 피자샌드위치..	0.151	948
긴긴추석연휴!! 메뉴가 걱정되세요? 매운갈비찜, 매운닭다리찜	0.150	604
전자렌지를 이용한 아주 맛있는 피자~ 떡베이컨말이피자	0.149	671

5. 결론

본 논문에서는 행렬 분해 기법의 하나인 SVD를 이용하여 블로그 공간을 분석하였다. SVD는 블로그 공간에서 커뮤니티들을 추출했고, 각 커뮤니티에서 영향력 높은 블로그들와 포스트들을 발견했다. 실제 블로그 공간을 분석함으로써 SVD가 블로그 공간을 분석하는데 유용하다는 것을 보였다.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단(No. 2008-0061006)의 지원을 받아 수행되었으며, 또한 지식경제부 및 정보통신산업진흥원의 대학 IT 연구센터 지원사업(NIPA-2010-(C1090-1011-0009))의 부분적인 지원을 받았습니다.

참고문헌

[1] R. Kumar, J. Novak, and A Tomkins, "Structure and Evolution of Online Social Networks," In *Proc. of Int'l. Conf. on Knowledge Discovery and Data*, KDD, pp. 611-617, 2006.

[2] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph Evolution: Densification and Shrinking Diameters," *ACM Trans. on Knowledge Discovery from Data*, ACM TKDD, Vol. 1, No. 1, Article 2, 2007.

[3] F. Korn, H. Jagadish, and C. Faloutsos, "Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences," In *Proc. of the 1997 ACM Special Interest Group on Management of Data Int'l. Conf. on Management of data*, ACM SIGMOD, pp. 289-300, 1997.

[4] S. Yoon, et al., "Extraction of a Latent Blog Community Based on Subject," In *Proc. of ACM Conf. on Information and Knowledge Management, ACM CIKM*, pp. 1529-1532, 2009.